General Comments

The authors present a study to investigate the impact of different domain sizes, vertical resolution, nesting ratios and spin-up time on a heavy precipitation event over Beijing. The simulations were forced by ERA-Interim reanalysis data available on 0.75° resolution in six hourly intervals. The different experiments were performed using three domains with a two-way nesting approach and the innermost domain centered on Beijing. Sub-daily precipitation of the second domain was verified against gridded precipitation observations from the China Meteorological Center. In addition, the precipitable water content of the WRF simulations was validated with ERA-Interim reanalysis data as a proxy for the maximum possible precipitation.

In a general sense, this type of experiments is of great relevance for flash flood forecasting and early warning systems. However, in the current experimental setup, I see several critical points preventing the traceability of the results.

1) The authors apply 2-way nesting from the outer to the inner domains. This means that precipitation patterns of the 3rd domain (which is not analyzed in your study) are reflected in the second domain. This actually means that you verify the precipitation from domain three mixed with terrain and land use data from domain two. Why did you verify domain two instead of domain three in your study? When looking at Fig. 1, Beijing has complex terrain which is not accurately represented at 13.5 km resolution.
   Also with a 2-way nesting approach, you do not balance any kind of model physics with respect to the lateral boundary conditions. In a 2-way nesting approach, the fine grid resolution replaces the coarser scale resolution over the area of domain three.

2) The authors decided to use ERA-Interim reanalysis data to initialize their model simulations. As mentioned on page four, the resolution is 0.75°. I am not sure if such a coarse resolution is able to provide reasonable initial conditions, especially when focusing on sub-daily rainfall. Although you mention that a small domain may benefit from the lateral boundary conditions, I doubt that such a small outer domain of effectively 30*30 grid cells (due to boundaries of at least 5 cells in each direction) is sufficient here. This is also mentioned on page five in your manuscript. If you carefully checked the WRF webpage, you may have noticed that least 100*100 cells are recommended for every domain.

3) It is also not clear how the WRF model levels are distributed in your simulation. From table 1, I only see that you used 29 levels up to 50 hPa. If a constant grid spacing of 1 km is applied, your model simulations will fail because processes in the PBL are not at all resolved. If you are in the middle troposphere, this spacing can be sufficient. Also the WRF tutorial and website suggest a vertical grid spacing of **less** than 1km. If you look at the user guide (e.g. http://www2.mmm.ucar.edu/wrf/users/docs/user_guide_V3.8/users_guide_chap5.htm#examples) you will see that at least ~40 levels are recommended when the model top is

set to 50 hPa. As your outer domain gets enlarged towards the pole, how did you deal with the map factors?

Choosing an adaptive time step may save computation time but is not the best way for scientific experiments (see WRF webpage).

4) The rescaling of the error measures may lead to a misinterpretation. In case of POD, how did you chose the factor 0.115? This is not clear from the manuscript. Is the maximum RMSE used for each individual time step or is it calculated from an average over all the time steps? It is also hard to believe that the POD remains constant, independent whether you start one week or 12 hours before the event? Also, what is the precipitation threshold used to calculate POD? Is it 0.1mm? Usually, POD is applied for different thresholds.

   How did you match both grids together? Did you use CDO, NCO, or NCL for this? It seems that you applied a $1/R^2$ approach to remap the CMC precipitation observations to the WRF grid. What is the radius of influence in this case? This can strongly determine the resulting field, especially in case of heavy and localized precipitation. Have the integrated water vapor fields been handled in the same way?

It would be very useful, if the authors provide horizontal plots of the precipitation patterns to substantiate the results. The applied scores do not necessarily tell if the precipitation is simulated spatially correctly.

In my opinion, a lot of important information is missing here and I also see deficiencies in the experimental setup. Therefore my recommendation is to reject the manuscript.

At the same time I encourage the authors to consider my suggestions and to resubmit a revised and updated version of the manuscript in the future. Before considering a resubmission, I strongly suggest that a native English speaker reads through the manuscript.